Useful things to know about probability theory

Basic ideas behind probability theory

When we approach probability from a theoretical perspective, we generally start by defining a set Ω of elementary events (sometimes also referred to as "outcomes"). For example, if we're modeling the throws of a die, it would be sensible to have Ω contain six elementary events, one for each face of the die. In general, you can think of elementary events as the things that can "happen" at the most basic level.

The next step is to define a set $\mathcal{F}$ of subsets of Ω. We call an element of $\mathcal{F}$ an "event" (not to be confused with the elementary events in Ω – an event is a set of elementary events). When Ω is a finite set, $\mathcal{F}$ contains *all* the subsets of Ω; that is, it is the power set of Ω. (Things get more complicated if Ω is infinite; generally, $\mathcal{F}$ is what is called a σ-algebra of Ω.)

We now get to the formal idea of probability. A probability P is essentially a function (technically, a measure) that takes events in $\mathcal{F}$ and assigns them a value in the interval $[0, 1]$. There are a few rules (axioms) that P needs to follow in order to match our intuitions about what probability is and to be mathematically useful:

1. For every $A \in \mathcal{F}$, $0 \leq P(A) \leq 1$.
2. $P(\Omega) = 1$.
3. The probability of the union of a finite or countable number of pairwise-disjoint events is equal to the sum of the probabilities of these events. That is, given a countable collection of pairwise-disjoint events $\{A_k : k = 1, 2, 3, \dots\}$,
$$P\left(\bigcup_k A_k\right) = \sum_k P(A_k).$$

These axioms can be used to derive a number of useful theorems. (For example, can you use these axioms to prove that $P(\emptyset) = 0$?) However, before going deeper into the theory, we should pause and clarify what all this means on a practical level.

Going back to the die example, let's suppose that we decide on $\Omega = \{1, 2, 3, 4, 5, 6\}$ to represent our elementary events. The set of all possible events $\mathcal{F}$ contains all of the $2^6 = 64$ possible subsets of Ω (including both the empty set and the set Ω itself). For any $A \in \mathcal{F}$, we can think of $P(A)$ as the probability that one of the elementary events in A happens. Thus, $P(\{1\})$ is the probability of rolling a 1, and $P(\{1, 3, 5\})$ is the probability of rolling an odd number.

Given $A, B \in \mathcal{F}$, we write the probability of *both* A and B occurring as $P(A \cap B)$. We use the set intersection operator $\cap$ here because we are literally asking, "What is the probability that one of the elementary events that is in both A and

B happens?" For example, the probability that we both roll an even number and a number greater than 3 is

$$P(\{2,4,6\} \cap \{4,5,6\}) = P(\{4,6\}).$$

Similarly, the probability of *either* A or B (or both) is $P(A \cup B)$.

When the set of elementary events is finite, one common way to define a valid probability function is assume that each elementary event is equally likely so that, $\forall A \in \mathcal{F}$,

$$P(A) = \frac{|A|}{|\Omega|}$$

The interpretation of this is that the probability of an event A is the number of elementary events in A divided by the *total* number of elementary events. This should make intuitive sense and is useful for making probability calculations in a lot of toy examples. Before we move on to explore the idea of a random variable, let's take a moment to consider one such example:

Suppose that we get bored with the single die we were using before and decide that rolling *five* dice all at once would be more fun. Let's see if we can apply the above ideas to figure out what that a given roll will have all five dice give the same number.

The first thing we need to do is figure out is what the set of elementary events looks like. Well, we can think of each possible outcome as a collection of five numbers, each number selected from the set $\{1,2,3,4,5,6\}$. For example, $(1,1,1,1,2)$ would represent the outcome that the first four dice we roll come up as 1, and the last die comes up as 2. Thus our set of all elementary events is $\Omega = \{1,2,3,4,5,6\}^5$ (the set of all ordered pairs with five coordinates, each from 1 to 5), and $|\Omega| = 6^5$.

Let's call A the event that all the dice come up as the same number. How many elementary events are there in A? Well, using our notation from before, it's pretty clear that the elementary events in A are $(1,1,1,1,1)$, $(2,2,2,2,2)$, $(3,3,3,3,3)$, $(4,4,4,4,4)$, $(5,5,5,5,5)$, and $(6,6,6,6,6)$. So $|A| = 6$.

If we assume that every possible roll of the dice is equally likely, we then have

$$P(A) = \frac{|A|}{|\Omega|} = \frac{6}{6^5} = \frac{1}{1296}$$

Generally, it's not necessary to be this explicit about defining the set of elementary events and all that when working through this sort of problem; however, keeping in mind the underlying mathematical ideas is helpful for fighting the propensity to just memorize formulas and procedures without really knowing what's going on.

Random Variables

The idea of a random variable seems to be rather befuddling for most students of statistics and probability (this was, at least, a confusing concept for me to get my head around at first), which is a shame, because things will make loads more sense if you can get a good grasp of what a random variable is.

> A *random variable* is a function that assigns a value (typically, a real number) to each elementary event in a set Ω of elementary events.

Take a moment to think about that. A random variable, at the most basic level, is function that maps elementary events to real numbers.

Let's define a random variable $X : \Omega \to \mathbb{R}$ and use to see why random variables are useful.

One thing to notice is that the inverse image under X of any subset of the real numbers is an event in $\mathcal{F}$. Specifically, for a subset $I \subseteq \mathbb{R}$, the reverse image of I under X is $\{\omega \in \Omega : X(\omega) \in I\}$. Since each event in $\mathcal{F}$ is associated with some probability, it is thus sensible to talk about the probability that X takes on some value in a given set. It is common to write something of the form

$$P(X \in I)$$

which is basically shorthand for

$$P(\{\omega \in \Omega : X(\omega) \in I\}).$$

Keep in mind that it is the underlying *events* that have probabilities. It is common to talk about the probability that a random variable equals some number or is in some interval, but what we're really talking about when we do that is the *probability of an event* that maps to that number/interval when plugged into the random variable.

Let's consider a couple of simple examples of random variables:

A discrete random variable

Suppose we have a bag that contains three balls: a red ball, a blue ball, and a royal ball. If we want to model the probability of randomly pulling out each ball, we could use the set of elementary events

$$\Omega = \{\text{red}, \text{blue}, \text{royal}\}.$$

Assuming that each ball is equally likely to be drawn, we can say things like

$$P(\{\text{royal}\}) = \frac{1}{3}$$

and

$$P(\{\text{red}, \text{blue}\}) = \frac{2}{3}.$$

Once we get bored with the idea that there's a 1/3 probability of a royal ball, we can make things more exciting by coming up with a function X that assigns some number to each of our elementary events. Let's say that $X(\text{red}) = 1$, $X(\text{blue}) = 2$, and $X(\text{royal}) = \pi$. What fun! Now we can say things like

$$P(X = 2) = \frac{1}{3}$$

and

$$P(X \in \mathbb{Q}) = \frac{2}{3}.$$

Remember that when we say that $P(X \in \mathbb{Q}) = 2/3$, what we're really saying is that the probability of some elementary event occurring that, when plugged into X, gives us a rational number is 2/3.

X is here an example of a *discrete random variable.* Discrete random variables take on a finite or countable number of values with nonzero probability. (Here X takes on just three values: 1, 2, and π.)

Note that whenever X is a discrete random variable, and $S \subset \mathbb{R}$ is the set of values that X takes with nonzero probability, $\sum_{x \in S} P(X = x) = 1$.

It is common to define a function $f_X : \mathbb{R} \to [0, 1]$ as $f_X(x) = P(X = x)$; we call f_X the probability function of X.

A continuous random variable

We say that a random variable X is a *continuous random variable* if there is a function $f_X : \mathbb{R} \to [0, \infty)$ such that for all $x \in \mathbb{R}$,

$$P(X < x) = \int_{-\infty}^{x} f_X(t)dt$$

We call f_X the *probability density function* of X. Pay attention to the fact that the probability density function of a continuous random variable is analogous to, but different from, the probability function of a discrete random variable: for a continuous random variable it is *not* the case that $f_X(x) = P(X = x)$. In fact, if X is a continuous random variable, then $P(X = x) = 0$ for all x. (Can you show why?)

In the mythical land of Nenie, the zany King Neniu will occasionally throw royal balls that he selects according to some random process from the set of all possible royal balls. How might we model King Neniu's royal balls using the ideas we've covered up to this point? We consider that there are uncountably-many possible royal balls, since certain aspects of the ball (such as the size of the ball room, the exact time it starts, or the amount of time it lasts) can be continuously modified. We can consider each possible royal ball to be an elementary event in a set Ω of all elementary events. Notice that it doesn't make sense to ask what the probability of any specific royal ball is – since there are an uncountable number

of royal balls to choose from, we could randomly select from Ω forever and *never* get the specific ball we're looking for. (This is why probabilities are defined on elements of a σ-algebra $\mathcal{F}$ of Ω rather than on Ω itself.) It *does* however make sense to ask things like, "What is the probability that a randomly selected ball will last between 4 and 5 hours?" We can define a random variable X to answer exactly that sort of question.

Let X be a function that maps each royal ball in Ω to the positive real number corresponding to the exact length of time (in units of hours) that the ball lasts. We can then interpret statements like "$P(X < 6)$" as meaning "What is the probability that a randomly selected royal ball lasts less than six hours?" Alternately, we could express that as

$$P(\{\omega \in \Omega : \omega \text{ lasts less than 6 hours}\}).$$

The random variable X is essentially a shorthand for that sort of statement. Since X takes an uncountable number of nonzero values (i.e. there are uncountably many possible lengths of time that a royal ball can last), it is not a discrete random variable. However, if the underlying probabilities behave "nicely," it could be a continuous random variable. For example, if balls with lengths between 1 and 8 hours are equally (uniformly) likely, and balls with lengths less than 1 hour or more than 8 hours never happen, then

$$f_X(x) = \begin{cases} 1/8, & x \in [1, 8] \\ 0, & \text{otherwise} \end{cases}$$

would be a probability density function for X.

When we use random variables in a practical setting, we typically work on the abstract level without explicitly acknowledging what the underlying events are and how they are mapped to the real numbers. But as I stated before, it never hurts to keep in mind the mathematical ideas behind what you're doing.

Parameters & statistics

A "parameter" is a characteristic of a random variable (and by extension, the population that that random variable is intended to model). The most-commonly considered parameters are the expected value (or mean) and the variance.

If X is a discrete random variable, its expected value $E[X]$ (commonly written as μ) is defined as

$$E[X] = \sum_x x f_X(x).$$

Analogously, if X is a continuous random variable, its expected value is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x).$$

The nth moment of a discrete random variable X is

$$E[X^n] = \sum_x x^n f_X(x),$$

and, in the continuous case,

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x).$$

The expected value of X is its first moment.

The variance of X, $Var(X)$ (commonly written as σ^2), is the difference of the second moment of X and the square of the expected value of X:

$$Var(X) = E[X^2] - E[X]^2$$

Note that this is equivalent to $Var(X) = E[(X - E[X])^2]$; intuitively, the variance is a measure of how much values tend to differ from the mean. The standard deviation σ is the square root of the variance.

In statistical theory, a "statistic" can be understood in either of two senses:

1. A statistic[1] is a *function* that takes a *sample* of observed values and maps them to a real number. Thus, a statistic[1] $s : \mathbb{R}^n \to \mathbb{R}$ would take values $x_1, x_2, \ldots x_n$ corresponding to observed random outcomes mapped to $\mathbb{R}$ by some random variables $X_1, X_2, \ldots X_3$. This is typically what people are talking about when they use the word "statistic."
2. A statistic[2] is a *random variable* defined as a composition of other random variables. That is, given a statistic[1] $s : \mathbb{R}^n \to \mathbb{R}$ and random variables $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$, we can create a statistic[2] $S : \Omega \to \mathbb{R}$ as a function of $X_1, X_2, \ldots, X_n$:

$$S(\omega) = s(X_1(\omega), X_2(\omega), \ldots, X_3(\omega))$$

You can think of a statistic[2] as the distribution from which a statistic[1] is drawn.

In what follows, let $X_1, X_2, \ldots, X_n$ be identically distributed random variables with mean μ and variance σ^2.

One common statistic[2] is the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

It's not too difficult to show that $E[\overline{X}] = \mu$ and $Var(\overline{X}) = \sigma^2/n$. This means that we can use the average of some observations $x_1, x_2, \ldots, x_n$ to get an unbiased estimate of the parameter μ whose reliability gets better as n increases.

Another statistic[2] with which you are likely familiar is the sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

It turns out that

$$E[S^2] = \frac{n-1}{n}\sigma^2$$

meaning that S^2 as an estimator for σ^2 is biased by a factor of $(n-1)/n$. We can correct this by instead using the *unbiased* sample variance

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Finding statistics that are unbiased estimates of population parameters is an important part of statistical analysis.